DOCUMENT RESUME

ED 269 436 TM 860 267

AUTHOR Jaeger, Richard M.; Busch, John C.
TITLE The Use and Effect of Caution Indices in Detecting Aberrant Patterns of Standard-Setting Recommendations.
PUB DATE Apr 86
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Error of Measurement; Evaluation Methods; High Schools; Interrater Reliability; Item Analysis; *Judges; *Judgment Analysis Technique; Measurement Techniques; Minimum Competency Testing; Response Style (Tests); Scores; Scoring; *Standards; Statistical Studies; Testing Problems
IDENTIFIERS *Iterative Methods; *Modified Caution Index; North Carolina High School Competency Tests

ABSTRACT

This study explores the use of the modified caution index (MCI) for identifying judges whose patterns of recommendations suggest that their judgments might be based on incomplete information, flawed reasoning, or inattention to their standard-setting tasks. It also examines the effect on test standards and passing rates when the test standards of judges whose recommendations appear aberrant are eliminated. Data for this study were collected in a project to establish standards on the North Carolina High School Competency Tests. In that study, judges provided recommendations on an appropriate standard for a 120-item reading test (n=83) and for a 120-item math test (n=84). Matrices of item-by-item recommendations provided by judges were ordered by total recommended test standard, and by number of item endorsements. Once MCI had been computed for each judge, distributions of MCI were constructed for each subtest, and the effects of eliminating judges with MCI greater than selected threshold values were determined. Results indicate that screening judges on the basis of their MCI values can make a substantial difference in the resulting test standard and in the resulting failure rate on a high school competency test. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED269436

The Use and Effect of Caution Indices in Detecting Aberrant Patterns of Standard-Setting Recommendations

Richard M. Jaeger and John C. Busch

University of North Carolina at Greensboro

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R M. Jaeger

J. C. Busch

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 860 267

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1986.

The Use and Effect of Caution Indices in Detecting Aberrant Patterns of Standard-Setting Recommendations

Richard M. Jaeger and John C. Busch
University of North Carolina at Greensboro

Since the late 1960's, various indices have been proposed for use in detecting unusual patterns of examinee response to sets of dichotomously scored test items. These indices include the personal biserial correlation (Donlon & Fischer, 1968), the caution index (Sato, 1975, as translated and reported by Tatsuoka, 1978), the U' index (van der Flier, 1977), the personal point-biserial index (Brennan, 1980), the agreement index and the disagreement index (Brennan & Kane, 1980), the norm-conformity index (Tatsuoka & Tatsuoka, 1980), and the modified caution index (Harnisch and Linn, 1981). Harnisch and Linn (1981) describe the algebraic relationships among these indices and report their intercorrelations for samples of more than 2000 fourth-grade students who completed achievement tests in reading and mathematics.

The objectives of this study were twofold. First, to examine the applicability of the modified caution index to analysis of the item recommendations provided by judges when they engaged in a judgmental procedure for setting a standard on a criterion-referenced test. And second, to investigate the effects on ultimate passing scores and examinees' passing rates when the test standards recommended by judges whose patterns of item recommendations were flagged as aberrant by the modified caution index were eliminated. The modified caution index was

the only index of its type selected for this investigation because it was found by Harnisch and Linn to be least highly correlated with examinees' total test scores.

Rationale

Widely-used judgmental procedures for setting standards on criterion-referenced tests (Nedelsky, 1954; Angoff,1971; Ebel, 1972; Jaeger, 1982) presume that judges are equally qualified and precise in providing their recommendations. Yet several empirical studies have shown that judges differ widely in the standards they recommend and in their item-by-item recommendations, regardless of the standard-setting method they use (c.f., Cross, et al., 1984). Available data (Jaeger and Busch, 1984) suggest that judges who recommend test standards are not equally informed of the tasks they are to complete nor are they equally confident of their abilities to complete standard-setting tasks. Systematic procedures for identifying such judges have not been widely applied, perhaps due to limited information on their effects and the complexity of methods that have been proposed in the past (van der Linden, 1982).

This study explores the use of the modified caution index for identifying judges whose patterns of recommendations suggest that their judgments might be based on incomplete information, flawed reasoning, or inattention to their standard-setting tasks. It also examines the effect on test standards and passing rates when the test standards of judges whose recommendations appear aberrant are eliminated.

Indices for detecting aberrant patterns of examinees' responses to items that compose an objective test have been used for a variety of purposes in Japan (Harnisch, 1983), and have been applied in the United States by Harnisch (1983), Tatsuoka and Tatsuoka (1982), and Miller (1981) to such diverse problems as identifying students for whom the standard interpretation of an examination score is potentially biased, identifying schools that differ in instructional emphasis, and identifying classes in which instruction has not been effective. However, these procedures have not been applied to the problem of screening judges in standard-setting applications.

Method

Data Source

Data for this study were collected in a project to establish standards on the North Carolina High School Competency Tests (Jaeger, et al., 1980). In that study, 83 judges provided recommendations on an appropriate standard for a 120-item reading test, and 84 judges provided recommendations on an appropriate standard for a 120-item mathematics test. Non-overlapping samples of judges were used for each test. Each sample was composed of high school teachers, registered voters, high school counselors, and high school principals.

Judges used the iterative, structured standard-setting procedure described by Jaeger (1982). In that procedure, each judge provided a "Yes/No" response to the question "Should every regular high school graduate in North Carolina be able to answer this item correctly?" for each test item. Judgments were coded

in dichotomous "1/0" form, and each judge's recommended test standard was defined as the sum of his/her coded responses. The data used in this study consisted of the coded item judgments recorded for the third (final) judgment session in the iterative standard-setting procedure (See Jaeger, 1982 for details).

Analytic Procedures

Consistent with procedures suggested by Harnisch (1983), matrices of item-by-item recommendations provided by judges for the reading test and the mathematics test during final judgment sessions were ordered by total recommended test standard, and by number of item endorsements, to form S-P charts.

The modified caution index (Harnisch & Linn, 1981) was then computed for each judge and for each test item in the reading and mathematics tests. The index was defined as follows:

$$C_i^* = \frac{\sum_{j=1}^{n_{i.}} (1 - u_{ij}) n_{.j} - \sum_{j=n_{i.}+1}^{J} u_{ij} n_{.j}}{\sum_{j=1}^{n_{i.}} n_{.j} - \sum_{j=J-(n_{i.}-1)}^{J} n_{.j}}$$

where

$i = 1,2, \ldots, I$ is the index for judges

$j = 1,2, \ldots, J$ is the index for test items

$$u_i = \begin{cases} 1 \text{ if judge i responds "Yes" to the question for item j} \\ 0 \text{ if judge i responds "No" to the question for item j} \end{cases}$$

$n_{i.}$ = the test standard recommended by the i^{th} judge

$n_{.j}$ = total number of "Yes" responses for the j^{th} item

Indices were computed by using Mainframe Version 2.0 of the Student Problem Package software (Harnisch, Kuo & Torres, 1983).

Once modified caution indices had been computed for each judge, distributions of modified caution indices were constructed for each subtest. These distributions were then systematically trimmed, to determine the effects of eliminating judges with modified caution indices greater than selected threshold values on (1) the numbers of judges whose judgments would be retained, (2) the passing scores that would result when these judges' recommendations were eliminated, and (3) associated changes in examinee passing rates.

Levels of association between the modified caution index and recommended test standards were also examined by computing Pearson correlation coefficients and through appropriate graphical procedures.

Results and Discussion

Behavior of the Modified Caution Index

Distributions of modified caution indices for judges rating items on the Reading Subtest and the Mathematics Subtest of the North Carolina High School Competency Tests are shown in Tables 1 and 2, respectively. For the Reading Subtest, modified caution indices ranged between 0.0550 and 0.6215, with a mean of 0.3257.

Insert Tables 1 and 2 About Here

For the Mathematics Subtest, modified caution indices ranged between 0.0440 and 0.6970, with a mean of 0.3031. The

distribution for the Mathematics Subtest showed slight positive skewness.

The modified caution index is scaled so that a value of zero indicates complete correspondence between the rank order of coded item recommendations assigned by an individual judge and the sum of the coded recommendations assigned by all judges. That is, a judge whose recommendations produced a modified caution index of zero would have answered "Yes" only for items having a higher total number of "Yes" judgments than would all items for which (s)he answered "No." A judge's recommendations would produce a modified caution index of 1.0 only if his/her recommendations were totally inconsistent with those of the entire sample of judges, in the sense that the judge answered "No" only for items having a higher total number of "Yes" judgements than would all items for which (s)he answered "Yes."

Because the modified caution index has not been applied previously to standard-setting recommendations, it is difficult to interpret the magnitudes of the index values reported here. In several recent papers, Harnisch & Linn (1980; 1981; 1982) have suggested that modified caution indices larger than 0.30 be labeled "large" and indicate patterns of examinee response that warrant further investigation. However, it is not at all clear that the same rule of thumb applies to patterns of item recommendations assigned by judges who are participating in a standard-setting exercise. The mean modified caution indices reported in Tables 1 and 2 are substantially larger than those reported in the Harnisch and Linn studies. In their analyses of mathematics items administered to 13-year-olds in the National

Assessment of Educational Progress (1982), Harnisch and Linn found a mean modified caution index of 0.18. Although their data tables do not permit computation of the quartiles of the distribution of modified caution indices, it is almost certain that the third quartile of their distribution is smaller than the median of ours. In their 1981 study, Harnisch and Linn report modified caution indices ranging from 0.18 to 0.32 for a reading test administered to fourth-graders in Illinois, and indices ranging from 0.13 to 0.26 for a mathematics test administered to the same students. In the context of these index distributions, an index value of 0.30 is clearly large. If larger index values typically are to be expected when standard-setting data are analyzed, this definition of a "large" index value will have to be modified accordingly.

When applied to traditional achievement test data, the modified caution index has been found to be moderately correlated with examinees' total test scores. In at least one study (Harnisch & Linn, 1982) the index was found to be negatively correlated with total test score. In this study, the modified caution index was found to be positively correlated with judges' recommended test standard, with the correlation being quite small for the Reading Subtest (0.158) and moderate (0.441) for the Mathematics Subtest. Patterns of relationship between the modified caution index and judges' recommended test standard are illustrated in Figures 1 and 2. Although Figure 1 shows a slight association between large values of the modified caution index

---

Insert Figures 1 and 2 About Here

---

and large values of recommended test score on the Reading Subtest, the relationship is quite modest. One would not therefore expect the elimination of recommendations of judges whose patterns of judgment produced large modified caution indices to have a large effect on the mean recommended test standard for the Reading Subtest. Subsequent analyses confirmed this expectation.

Figure 2 shows a far more definite pattern of relationship between values of the modified index and judges' recommended standards for the Mathematics Subtest. The relationship appears to be somewhat non-linear, with near-uniform distributions of the modified caution index for judges whose recommended test standard were smaller than 108, and decidedly larger index values for most judges whose recommended test standards were larger than 108. We would therefore assume that elimination of the recommendations of judges with the largest modified caution indices would reduce the mean recommended standard on the Mathematics Subtest.

Effects on Test Standards

The effects on mean recommended test standards of eliminating the recommendations of judges whose modified caution index values were larger than specified thresholds are shown in Figures 3 and 4 for the Reading Subtest and the Mathematics Subtest, respectively.

Of the 83 judges who provided recommendations on the Reading Subtest, four recommended either zero or 120 as the test standard. Since modified caution indices cannot be computed

for judges whose recommendations are at either extreme, the recommendations of these judges were excluded from further analyses. The same procedure was used to trim the sample of judges who provided recommendations on the Mathematics Subtest from 84 to 79.

The mean test standard recommended by the remaining 79 judges who considered items on the Reading Subtest was 99.51. As can be seen in Figure 3, this mean recommended test standard was changed by less than one point by excluding judges whose modified caution indices exceeded values ranging from 0.60 to 0.40. When judges whose modified caution index exceeded 0.60 were excluded, the mean test standard recommended by the remaining 77 judges was 99.29. The sizes of remaining samples of judges and mean recommended standards on the Reading Subtest corresponding to other threshhold values are as follows:

Threshold for MCI	Sample Size	Mean Recomm. Standard
0.55	73	98.99
0.50	72	98.81
0.45	64	99.08
0.40	53	98.87
0.35	46	98.61
0.30	36	96.97
0.25	29	97.66
0.20	18	96.50
0.15	6	98.00

Insert Figure 3 About Here

Not until a threshold value of 0.30 was reached, at which point only 36 of 79 judges remained in the sample, was the resulting mean recommended test standard for the Reading Subtest changed by

more than one score point. These results suggest that judges whose pattern of recommendations were flagged as "aberrant" by the modified caution index differed only in their pattern of recommendations, and not in the number of test items for which they responded "Yes," from judges whose pattern or recommendations produced small values of the modified caution index. It is important to note that setting the threshold value of the modified caution index low enough to affect the mean recommended standard on the Reading Subtest by more than one point would require excluding the judgments of over half the original judgment panel. The consequent loss of judgment reliability would be substantial, and would make such a reduction unreasonable.

Items on the Mathematics Subtest were reviewed by 84 judges, of which five recommended standards of zero or 120. The mean standard recommended by the other 79 judges was 96.59. Since for the Mathematics Subtest, judges' modified caution index values were substantially correlated with their recommended standards, one would expect the mean recommended test standard to change appreciably as the recommendations of judges with large modified caution index values were eliminated. The graph in Figure 4 confirms this expectation.

Insert Figure 4 About Here

When the recommendations of judges whose modified caution indices exceeded 0.65 were eliminated, the mean standard recommended by the remaining 77 judges was 96.12. The sizes of

remaining samples of judges and mean recommended standards on the Mathematics Subtest corresponding to other threshold values are as follows:

Threshold for MCI	Sample Size	Mean Recomm. Standard
0.60	76	95.89
0.55	73	95.23
0.50	73	95.23
0.45	69	94.67
0.40	64	94.09
0.35	57	92.96
0.30	44	91.34
0.25	30	89.93
0.20	19	90.32

In contrast to the results reported above for the Reading Subtest, eliminating the recommendations of Mathematics Subtest judges whose modified caution indices exceeded a moderately large threshold value would have a noticeable effect on the mean recommended test standard. For example, by using a threshold value of 0.35, the mean recommended standard would be reduced by three points. The resulting test standard would be based on the judgments of almost three-fourths of the original judgment panel, with a negligible reduction in the reliability of the mean recommended test standard.

One potential benefit of screening judges with apparently aberrant patterns of recommendations would be a reduction in the standard error of the mean recommended test standard. One might reasonably expect that the test standards recommended by the remaining judges would be more homogeneous, once screening had taken place. Unfortunately, the results observed in this study do not confirm this expectation. The standard deviation of the test standards recommended by the entire sample of 79 judges who

considered items on the Reading Subtest was 10.18. The corresponding standard deviation for the 29 judges whose modified caution indices were less than 0.20 was 10.02. Results for the Mathematics Subtest were similar. The standard deviation of recommended test standards for the entire sample of 79 judges was 14.39. For the 57 judges whose modified caution indices were less than 0.35, the corresponding standard deviation was 13.41, and for the 19 judges whose modified caution indices were less than 0.20, the corresponding standard deviation was 15.26. There is no evidence in these data that judges with smaller modified caution index values are more homogeneous in their overall recommended test standards.

Effects on Examinee Failure Rates

Figures 5 and 6 illustrate the effects on examinee failure rates of screening judges whose modified caution indices exceeded selected threshold values. As would be expected from the negligible effects of screening on the mean standard recommended for the Reading Subtest, discussed above, the failure rate on the Reading Subtest is not appreciably affected by excluding the recommendations of judges whose modified caution indices exceed 0.30 (see Figure 5). Even with lower threshold values, the effects of screening judges on the examinee failure rate are inconsistent. The failure rate drops from 23 percent (without screening of judges) to 20 percent (when the recommendations of judges whose modified caution index exceeds 0.30 are eliminated), but then increases to 22 percent when the MCI threshold is further reduced to 0.25. The 19 percent failure rate associated

with a threshold value of 0.20 is based on a sample of 18 judges -- less than one-fourth of the original judgment panel for the Reading Subtest.

Insert Figures 5 and 6 About Here

Consistent with results discussed above, substantially greater effects of screening judges were found for examinee failure rates on the Mathematics Subtest (see Figure 6). When the recommendations of judges whose modified caution indices exceeded 0.45 were eliminated, the failure rate was reduced by three percent. Setting the MCI threshold at 0.35 resulted in a reduction in the examinee failure rate of five percent. Further reductions in the MCI threshold had monotonic effects on the examinee failure rate on the Mathematics Subtest, with a nine percent reduction resulting from an MCI threshold of 0.25 (based on the recommendations of a remaining sample of 30 judges).

Interpretation and Implications

When applied to conventional test data, the meaning of a large modified caution index is relatively clear. An examinee with such a value has responded in an unconventional manner to a set of test items. Harnisch and Linn (1982, p. 6) suggested that large MCI values might result from examinees having guessed the answers to test items, from their being careless, from their being highly anxious as they completed the test, from their having had unusual instructional histories (e.g., having received instruction in the topics of the more difficult test items, but

no instruction in the topics of test items that are less difficult), from their having atypical experiential backgrounds, from their having copied other examinees' answers to some test items, or from their misunderstanding of specific content contained in some of the less difficult test items. In making these interpretations, it is implicitly assumed that a majority of examinees would not be described similarly, since it is their responses to individual test items that determine the items' difficulty values. Or, put another way, if an examinee earns a high MCI value it is because his/her pattern of item responses differs from those of the majority, not because some test items are inherently more or less difficult than others.

Similarly, when a judge who participates in a standard-setting exercise exhibits a pattern of recommendations that produces a large MCI value, one can conclude that the judge's pattern of recommendations is different from those of the majority of judges. When computed in the context of Jaeger's (1982) standard-setting procedure, a large MCI value means that the judge recommended "Yes" on a lot of test items for which the majority recommended "No," and vice-versa. But what further interpretation can be made? The potential reasons for a large MCI value are not as evident when the index is applied to standard-setting data.

The complexity of Jaeger's standard-setting procedure is justified by his desire to inform judges of the implications of their recommendations, to provide them with several opportunities to reconsider their initial judgments, and to allow judges to consider the actual item-by-item performances of examinees who

will take the test for which they are recommending a standard. It might reasonably be argued that judges whose patterns of recommendations differ markedly from those of the majority are, to some degree, responding differently to the information they have received during the iterative standard-setting exercise. Whether it is reasonable to eliminate the recommendations of judges whose MCI values are large depends on the reasons such judges are responding differently. If judges whose recommendations lead to large MCI values have failed to understand the standard-setting task they are to perform, or if they have not attended to data on the implications of their recommended standards because they did not understand it, their recommendations are not reasoned, and should clearly be eliminated. Conversely, if judges whose recommendations produce large MCI values are no less informed and reasoning than are judges whose recommendations produce small MCI values, elimination of their recommendations is unwarranted.

Achieving consensus on an appropriate standard for a test is an admirable goal (certainly guaranteed through the use of a single judge) but it should not be pursued at the expense of fairly representing the population of judges whose recommendations are pertinent to the task of establishing a workable and equitable test standard. To eliminate the recommendations of some judges <u>just</u> because they differ from those of the majority is antithetical to the more fundamental goal of seeking the informed and reasoned judgments of one or more samples of judges who represent the population(s) of persons who have a legitimate stake in the outcome of the testing program

under study.

This argument suggests that procedures for screening judges be used cautiously and thoughtfully, if at all. Substantially more research -- on such questions as the relationships between judges' MCI values and their (1) confidence in their standard-setting recommendations, (2) understanding of the standard-setting task, (3) knowledge of the population for which test standards are being established, and (4) knowledge of the subject-matter of the test for which test standards are being established -- is needed before the use of the modified caution index, or any other procedure for screening judges, can be recommended.

It would seem more reasonable to eliminate the recommendations of some judges because they make poor judgments, rather than eliminating their recommendations because their pattern of recommendations is inconsistent with those of the majority. One approach to such screening would have judges make item-by-item recommendations for a test on which the item-performances of marginally competent examinees were known. The modified caution index could then be computed for each judge, using his/her "Yes/No" recommendations coded in "1/0" form, and the item performances of marginally competent examinees would compose the balance of the S-P chart. Judges would then earn large MCI values only because their recommendations were inconsistent with the performances of actual examinees; that is, because their judgments were incorrect.

The present study has shown that screening judges on the basis of their modified caution index values can make a substantial difference in the resulting test standard and in the

resulting failure rate on a high school competency test. Whether such screening should be employed routinely, and which screening method is most reasonable, are topics for additional research.

References

Angoff, W. H. (1971) Scales, norms, and equivalent scores. In R. L. Thorndike (ed.) Educational Measurement (2nd. ed.). Washington, D. C.: American Council on Education.

Cross, L. H., Impara, J. C., Frary, R. B. & Jaeger, R. M. (1984) A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-129.

Ebel, R. L. (1972) Essentials of Educational Measurement. Englewood Cliffs, N. J.: Prentice-Hall.

Harnisch, D. L. (1983) Item response patterns: Applications for educational practice. Journal of Educational Measurement, 20, 191-206.

Harnisch, D. L. & Linn, R. L. (1981) Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement,18, 133-146.

Harnisch, D. L. & Linn, R. L. (1982) Identification of aberrant response patterns in a test of mathematics. Paper presented at the Annual Meeting of the American Educational Research Association, New York City.

Harnisch, D. L., Kuo, S., & Torres, R. (1983) Student Problem Package User's Guide, Version 2.0. Urbana, IL: University of Illinois, Office of Educational Testing, Research and Service.

Jaeger, R. M. (1982) An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-475.

Jaeger, R. M. & Busch,J. C. (1984) A validation and standard-setting study of the General Knowledge and Communication Skills Tests of the National Teacher Examinations: Final Report. Greensboro, NC: Center for Educational Research and Evaluation, University of North Carolina at Greensboro.

Jaeger, R. M., Cole, J., Irwin, D. M. & Pratto, D. J. (1980) An iterative structured judgment process for setting passing scores on competency tests applied to the North Carolina High School Competency Tests in Reading and Mathematics. Greensboro, NC: Center for Educational Research and Evaluation, University of North Carolina at Greensboro.

Miller, M. D. (1981) Measuring between-group differences in instruction. Unpublished doctoral dissertation, University of California, Los Angeles.

Nedelsky, L. (1954) Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Sato, T. (1975) [The construction and interpretation of S-P tables.] Tokyo: Meiji Tosho.

Tatsuoka, M. M. (1978) Recent psychometric advances in Japan: Engineers grapple with educational measurement problems. Paper presented at the ONR Contractors Meeting on Individualized Measurement, Columbia, MO.

Tatsuoka, K. K. & Tatsuoka, M. M. (1982) Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 7, 215-231.

van der Flier, H. (1982) Deviant response patterns and comparability of test scores. Journal of Cross-Cultural Psychology, 13, 267-298.

van der Linden, W. J. (1982) A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 19, 295-308.

Table 1

Distribution of Modified Caution Indices

for

Judges Who Rated Items on the North Carolina

High School Competency Test in Reading

Stem-and-Leaf Display (Leaf Unit = 0.01)

Stem	Leaves
0	57899
1	4
1	667778889999
2	01112233344
2	5678899
3	0112222344
3	6777899
4	12222333444
4	56679999
5	2
5	5678
6	12

Minimum = 0.0550
Q_1 = 0.2138
Median = 0.3238
Q_3 = 0.4364
Maximum = 0.6215

Mean = 0.3257
S. Dev.= 0.1392

Sample Size = 79

Table 2

Distribution of Modified Caution Indices

for

Judges Who Rated Items on the North Carolina

High School Competency Test in Mathematics

Stem-and-Leaf Display (Leaf Unit = 0.01)

Stem	Leaves
0	4
0	
1	2334
1	55566677888899
2	00002223334
2	55666777888889
3	0001112223444
3	5888899
4	00022
4	6779
5	
5	588
6	2
6	69

Minimum = 0.0440
Q_1 = 0.2011
Median = 0.2845
Q_3 = 0.3857
Maximum = 0.6970

Mean = 0.3031
S. Dev.= 0.1316

Sample Size = 79

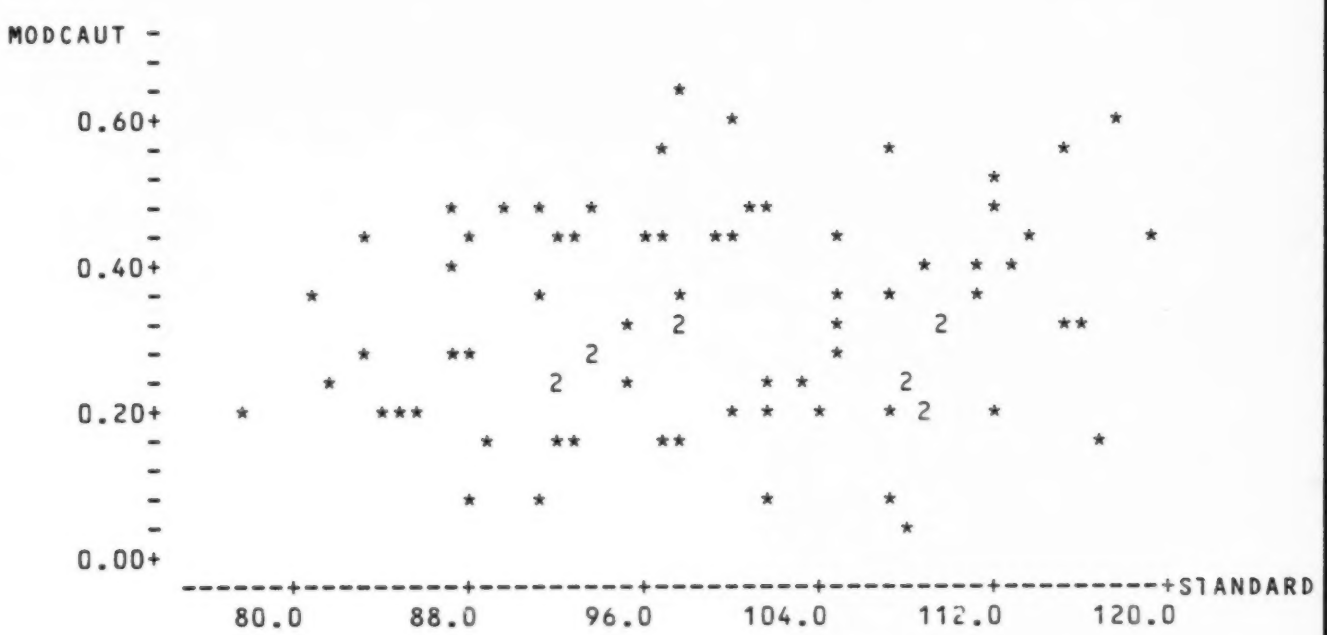

Figure 1. Plot of Modified Caution Index vs. Recommended Standard on the Reading Subtest of the North Carolina High School Competency Tests.

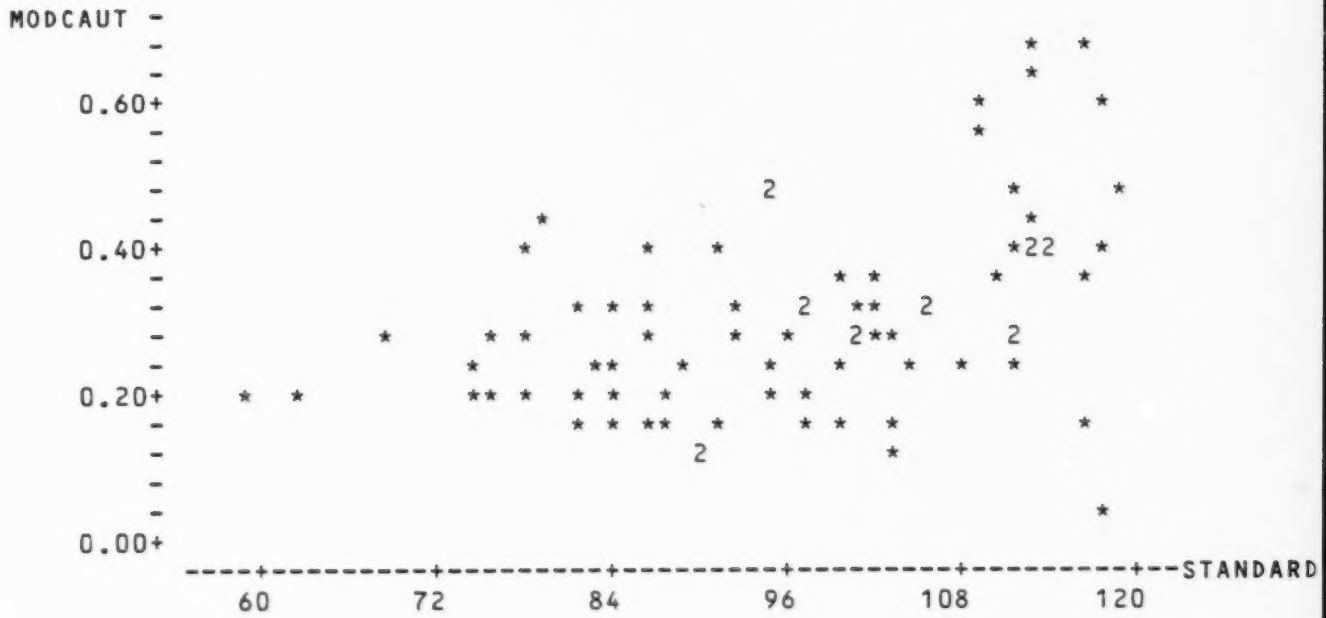

Figure 2. Plot of Modified Caution Index vs. Recommended Standard on the Mathematics Subtest of the North Carolina High School Competency Tests.

Figure 3. Mean Recommended Standard on the Reading Subtest of the North Carolina High School Competency Tests as a Function of the Maximum Modified Caution Index Retained.

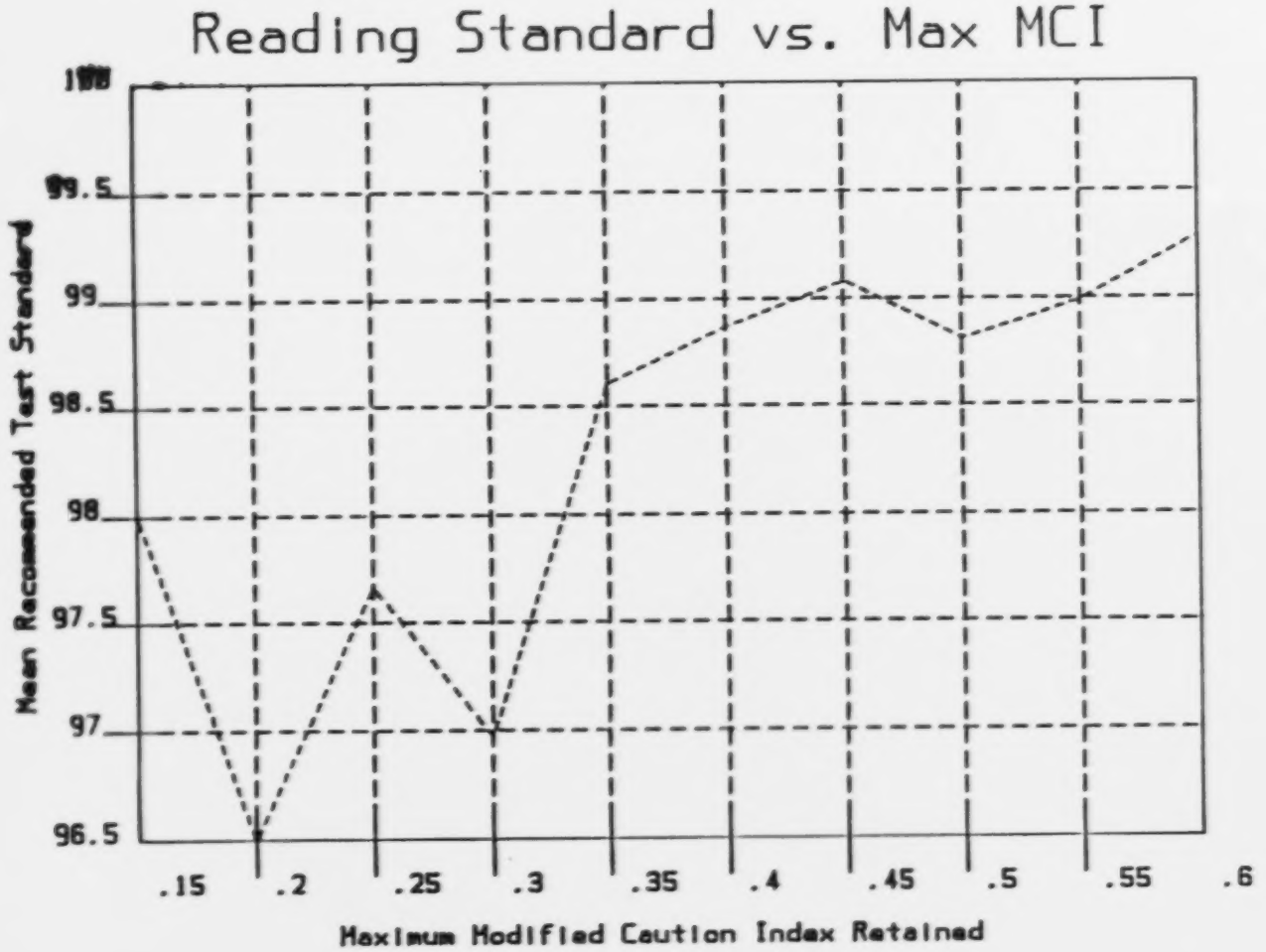

Figure 4. Mean Recommended Standard on the Mathematics Subtest of the North Carolina High School Competency Tests as a Function of the Maximum Modified Caution Index Retained.

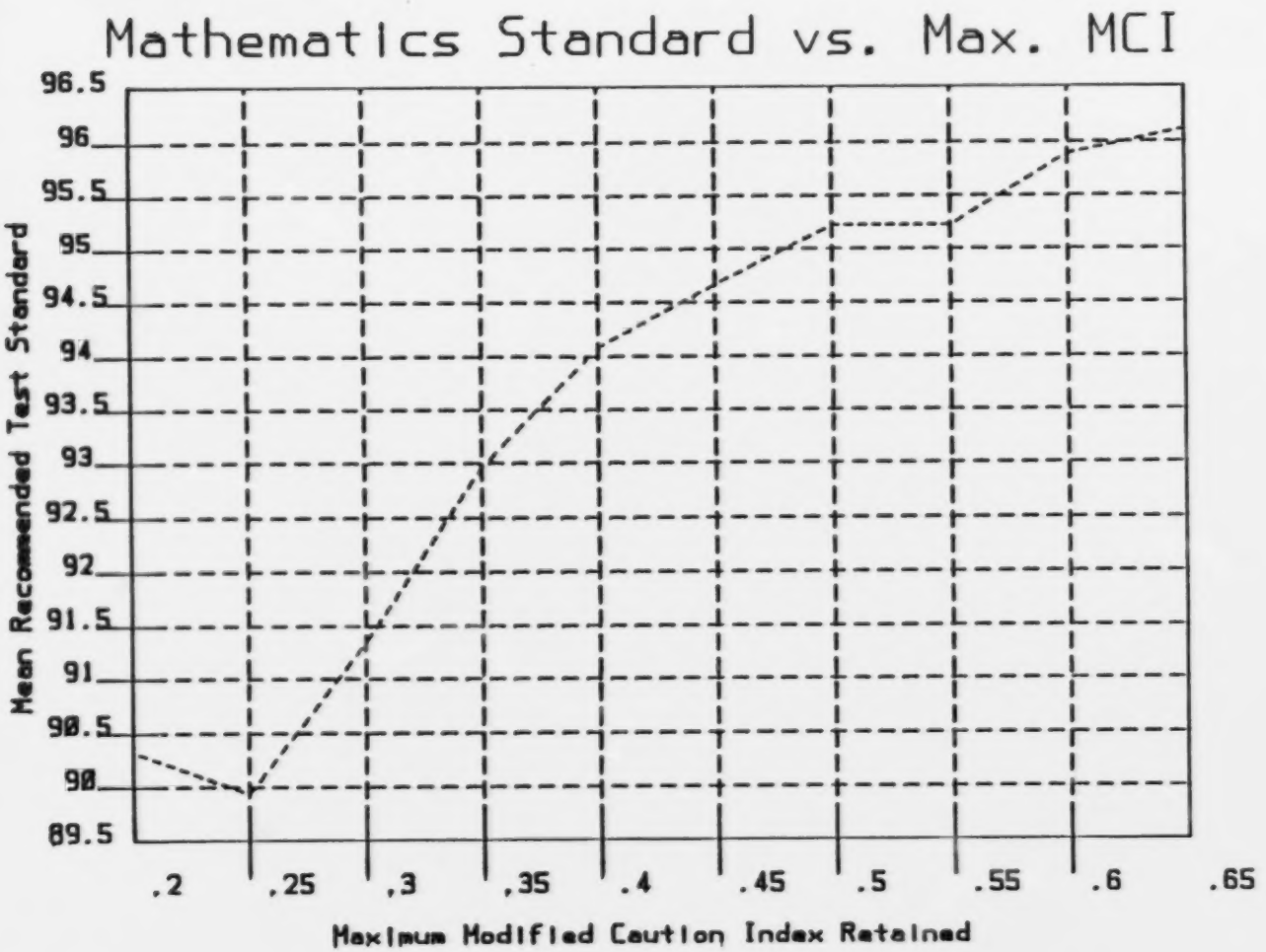

Figure 5. Percent of Examinees Who Would Fail the Reading Subtest of the North Carolina High School Competency Tests as a Function of the Maximum Modified Caution Index Retained.

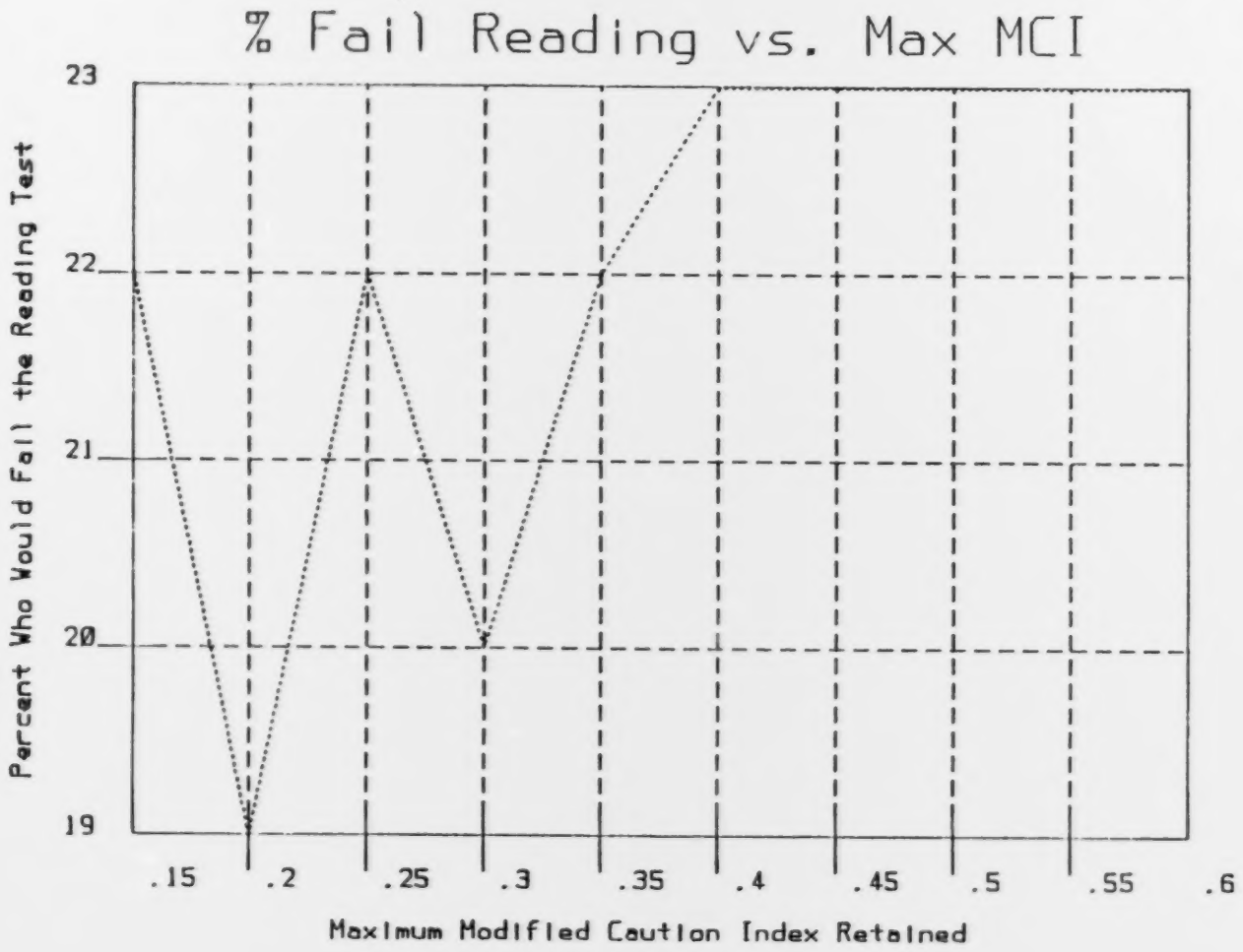

Figure 6. Percent of Examinees Who Would Fail the Mathematics Subtest of the North Carolina High School Competency Tests as a Function of the Maximum Modified Caution Index Retained.